Analysis of IMDB Data

We will analyze a subset of IMDB's actors, genres, movie actors, and movie ratings data. This dataset comes to us from Kaggle (https://www.kaggle.com/datasets/ashirwadsangwan/imdb-dataset) although we have taken steps to pull this data into a publis s3 bucket:

- s3://cis9760-lecture9-movieanalysis/name.basics.tsv ---> (actors)
- s3://cis9760-lecture9-movieanalysis/title.basics.tsv ---> (genres)
- s3://cis9760-lecture9-movieanalysis/title.principals.tsv ---> (movie actors)
- s3://cis9760-lecture9-movieanalysis/title.ratings.tsv ---> (movie ratings)

Content

name.basics.tsv.gz – Contains the following information for names:
nconst (string) - alphanumeric unique identifier of the name/person.
primaryName (string)– name by which the person is most often credited.
birthYear – in YYYY format.
deathYear – in YYYY format if applicable, else .
primaryProfession (array of strings)– the top-3 professions of the person.
knownForTitles (array of tconsts) – titles the person is known for.

title.basics.tsv.gz - Contains the following information for titles:
tconst (string) - alphanumeric unique identifier of the title.
titleType (string) – the type/format of the title (e.g. movie, short, tvseries, tvepisode, video, etc).
primaryTitle (string) – the more popular title / the title used by the filmmakers on promotional materials at the point of release.
originalTitle (string) - original title, in the original language.
isAdult (boolean) - 0: non-adult title; 1: adult title.
startYear (YYYY) – represents the release year of a title. In the case of TV Series, it is the series start year.
endYear (YYYY) – TV Series end year. for all other title types.
runtimeMinutes – primary runtime of the title, in minutes.
genres (string array) – includes up to three genres associated with the title.

title.principals.tsv – Contains the principal cast/crew for titles:
tconst (string) - alphanumeric unique identifier of the title.
ordering (integer) – a number to uniquely identify rows for a given titleId.
nconst (string) - alphanumeric unique identifier of the name/person.

category (string) - the category of job that person was in.
job (string) - the specific job title if applicable, else.
characters (string) - the name of the character played if applicable, else.

title.ratings.tsv.gz – Contains the IMDb rating and votes information for titles:
tconst (string) - alphanumeric unique identifier of the title.
averageRating – weighted average of all the individual user ratings.
numVotes - number of votes the title has received.

PART 1 - Installation and Initial Setup

Begin by installing the necessary libraries that you may need to conduct your analysis. At the very least, you must install pandas and matplotlib

In [2]:
```
%%info
```

Current session configs: `{'conf': {'spark.pyspark.python': 'python3', 'spark.pyspark.virtualenv.enabled': 'true', 'spark.pyspark.virtualenv.type': 'native', 'spark.pyspark.virtualenv.bin.path': '/usr/bin/virtualenv'}, 'kind': 'pyspark'}`

ID	YARN Application ID	Kind	State	Spark UI	Driver log	Current session?
5	application_1669654196452_0006	pyspark	idle	Link	Link	✓

Let's install the necessary packages here

In [3]:
```
sc.install_pypi_package("pandas==1.0.3")
sc.install_pypi_package("matplotlib==3.2.1")
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

```
Collecting pandas==1.0.3
  Using cached https://files.pythonhosted.org/packages/4a/6a/94b219b8ea0f2d580169e85e
d1edc0163743f55aaeca8a44c2e8fc1e344e/pandas-1.0.3-cp37-cp37m-manylinux1_x86_64.whl
Requirement already satisfied: pytz>=2017.2 in /usr/local/lib/python3.7/site-packages
(from pandas==1.0.3)
Requirement already satisfied: numpy>=1.13.3 in /usr/local/lib64/python3.7/site-packa
ges (from pandas==1.0.3)
Collecting python-dateutil>=2.6.1 (from pandas==1.0.3)
  Using cached https://files.pythonhosted.org/packages/36/7a/87837f39d0296e723bb9b62b
bb257d0355c7f6128853c78955f57342a56d/python_dateutil-2.8.2-py2.py3-none-any.whl
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (fr
om python-dateutil>=2.6.1->pandas==1.0.3)
Installing collected packages: python-dateutil, pandas
Successfully installed pandas-1.0.3 python-dateutil-2.8.2

Collecting matplotlib==3.2.1
  Using cached https://files.pythonhosted.org/packages/b2/c2/71fcf957710f3ba1f09088b3
5776a799ba7dd95f7c2b195ec800933b276b/matplotlib-3.2.1-cp37-cp37m-manylinux1_x86_64.wh
l
Requirement already satisfied: python-dateutil>=2.1 in /mnt/tmp/1669693905583-0/lib/p
ython3.7/site-packages (from matplotlib==3.2.1)
Collecting pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 (from matplotlib==3.2.1)
  Using cached https://files.pythonhosted.org/packages/6c/10/a7d0fa5baea8fe7b50f448ab
742f26f52b80bfca85ac2be9d35cdd9a3246/pyparsing-3.0.9-py3-none-any.whl
Collecting cycler>=0.10 (from matplotlib==3.2.1)
  Using cached https://files.pythonhosted.org/packages/5c/f9/695d6bedebd747e5eb0fe8fa
d57b72fdf25411273a39791cde838d5a8f51/cycler-0.11.0-py3-none-any.whl
Requirement already satisfied: numpy>=1.11 in /usr/local/lib64/python3.7/site-package
s (from matplotlib==3.2.1)
Collecting kiwisolver>=1.0.1 (from matplotlib==3.2.1)
  Using cached https://files.pythonhosted.org/packages/ab/8f/8dbe2d4efc4c0b08ec67d6ef
b7cc31fbfd688c80afad85f65980633b0d37/kiwisolver-1.4.4-cp37-cp37m-manylinux_2_5_x86_6
4.manylinux1_x86_64.whl
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (fr
om python-dateutil>=2.1->matplotlib==3.2.1)
Collecting typing-extensions; python_version < "3.8" (from kiwisolver>=1.0.1->matplot
lib==3.2.1)
  Using cached https://files.pythonhosted.org/packages/0b/8e/f1a0a5a76cfef77e1eb6004c
b49e5f8d72634da638420b9ea492ce8305e8/typing_extensions-4.4.0-py3-none-any.whl
Installing collected packages: pyparsing, cycler, typing-extensions, kiwisolver, matp
lotlib
Successfully installed cycler-0.11.0 kiwisolver-1.4.4 matplotlib-3.2.1 pyparsing-3.0.
9 typing-extensions-4.4.0
```

Now, import the installed packages from the previous block below.

In [4]:

```python
import pandas as pd
import numpy as np
import itertools
from pyspark.sql.functions import split, col, explode, approx_count_distinct, avg, col
import matplotlib.pyplot as plt
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

Loading Data

Load all data from S3 into a Spark dataframe object

In [5]:
```
actors = spark.read.csv('s3://cis9760-lecture9-movieanalysis/name.basics.tsv', sep=r'\
genres = spark.read.csv('s3://cis9760-lecture9-movieanalysis/title.basics.tsv', sep=r'
movie_actors = spark.read.csv('s3://cis9760-lecture9-movieanalysis/title.principals.ts
movie_ratings = spark.read.csv('s3://cis9760-lecture9-movieanalysis/title.ratings.tsv'
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

Actors

Display the schema below:

In [6]:
```
actors.printSchema()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
root
 |-- nconst: string (nullable = true)
 |-- primaryName: string (nullable = true)
 |-- birthYear: string (nullable = true)
 |-- deathYear: string (nullable = true)
 |-- primaryProfession: string (nullable = true)
 |-- knownForTitles: string (nullable = true)
```

Display the first 5 rows with the following columns:

- `primaryName`
- `birthYear`
- `deathYear`
- `knownForTitles`

In [7]:
```
actors.select("primaryName","birthYear","deathYear","knownForTitles").show(5, truncate
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+---------------+---------+---------+---------------------------------------+
|primaryName    |birthYear|deathYear|knownForTitles                         |
+---------------+---------+---------+---------------------------------------+
|Fred Astaire   |1899     |1987     |tt0050419,tt0053137,tt0072308,tt0043044|
|Lauren Bacall  |1924     |2014     |tt0071877,tt0117057,tt0038355,tt0037382|
|Brigitte Bardot|1934     |\N       |tt0054452,tt0049189,tt0059956,tt0057345|
|John Belushi   |1949     |1982     |tt0077975,tt0072562,tt0080455,tt0078723|
|Ingmar Bergman |1918     |2007     |tt0069467,tt0050976,tt0083922,tt0050986|
+---------------+---------+---------+---------------------------------------+
only showing top 5 rows
```

Genres

Display the first 10 rows with the following columns:

- `titleType`
- `primaryTitle`
- `genres`

In [8]:
```
genres.select("titleType","primaryTitle","genres").show(10, truncate=False)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+---------+-------------------------------------------+------------------------+
|titleType|primaryTitle                               |genres                  |
+---------+-------------------------------------------+------------------------+
|short    |Carmencita                                 |Documentary,Short       |
|short    |Le clown et ses chiens                     |Animation,Short         |
|short    |Pauvre Pierrot                             |Animation,Comedy,Romance|
|short    |Un bon bock                                |Animation,Short         |
|short    |Blacksmith Scene                           |Comedy,Short            |
|short    |Chinese Opium Den                          |Short                   |
|short    |Corbett and Courtney Before the Kinetograph|Short,Sport             |
|short    |Edison Kinetoscopic Record of a Sneeze     |Documentary,Short       |
|movie    |Miss Jerry                                 |Romance                 |
|short    |Exiting the Factory                        |Documentary,Short       |
+---------+-------------------------------------------+------------------------+
only showing top 10 rows
```

Display the unique categories below:

In [9]:
```
genres.select("titleType").distinct().show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+------------+
|   titleType|
+------------+
|    tvSeries|
|tvMiniSeries|
|       movie|
|   videoGame|
|   tvSpecial|
|       video|
|     tvMovie|
|   tvEpisode|
|     tvShort|
|       short|
+------------+
```

Display the schema below:

In [10]:
```
genres.printSchema()
```

```
VBox()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
root
 |-- tconst: string (nullable = true)
 |-- titleType: string (nullable = true)
 |-- primaryTitle: string (nullable = true)
 |-- originalTitle: string (nullable = true)
 |-- isAdult: string (nullable = true)
 |-- startYear: string (nullable = true)
 |-- endYear: string (nullable = true)
 |-- runtimeMinutes: string (nullable = true)
 |-- genres: string (nullable = true)
```

Movie Actors

Display the schema below:

In [11]:
```
movie_actors.printSchema()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
root
 |-- tconst: string (nullable = true)
 |-- ordering: string (nullable = true)
 |-- nconst: string (nullable = true)
 |-- category: string (nullable = true)
 |-- job: string (nullable = true)
 |-- characters: string (nullable = true)
```

Display the first 10 rows below

In [12]:
```
movie_actors.show(10, truncate=False)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+---------+--------+---------+---------------+-----------------------+-----------+
|tconst   |ordering|nconst   |category       |job                    |characters |
+---------+--------+---------+---------------+-----------------------+-----------+
|tt0000001|1       |nm1588970|self           |\N                     |["Herself"]|
|tt0000001|2       |nm0005690|director       |\N                     |\N         |
|tt0000001|3       |nm0374658|cinematographer|director of photography|\N         |
|tt0000002|1       |nm0721526|director       |\N                     |\N         |
|tt0000002|2       |nm1335271|composer       |\N                     |\N         |
|tt0000003|1       |nm0721526|director       |\N                     |\N         |
|tt0000003|2       |nm5442194|producer       |producer               |\N         |
|tt0000003|3       |nm1335271|composer       |\N                     |\N         |
|tt0000003|4       |nm5442200|editor         |\N                     |\N         |
|tt0000004|1       |nm0721526|director       |\N                     |\N         |
+---------+--------+---------+---------------+-----------------------+-----------+
only showing top 10 rows
```

Movie Ratings

Display the schema below:

```
In [13]: movie_ratings.printSchema()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
root
 |-- tconst: string (nullable = true)
 |-- averageRating: string (nullable = true)
 |-- numVotes: string (nullable = true)
```

Display the first 10 rows in a descending order by the number of votes

```
In [14]: movie_ratings.sort(movie_ratings.numVotes.desc()).show(10, truncate=False)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+---------+-------------+--------+
|tconst   |averageRating|numVotes|
+---------+-------------+--------+
|tt7430722|6.8          |9999    |
|tt4445154|8.1          |9997    |
|tt2229907|6.3          |9996    |
|tt0294097|8.0          |9994    |
|tt0264734|6.5          |9993    |
|tt2032572|5.2          |9991    |
|tt8860450|6.3          |9991    |
|tt3244036|8.3          |999     |
|tt1739480|6.9          |999     |
|tt1859607|5.3          |999     |
+---------+-------------+--------+
only showing top 10 rows
```

Overview of Data

Display the number of rows and columns in each dataFrame object.

```
In [15]: dflist=[actors,genres,movie_actors,movie_ratings]
tablelist=['Actors','Genres','Movie Actors','Movie Ratings']
for (df,tablename) in zip(dflist,tablelist):
    cols = len(df.columns)
    rows = df.count()
    print(f"Number of columns in {tablename} table: {cols}\nNumber of rows in {tablena
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

```
Number of columns in Actors table: 6
Number of rows in Actors table: 9706922

Number of columns in Genres table: 9
Number of rows in Genres table: 6321302

Number of columns in Movie Actors table: 6
Number of rows in Movie Actors table: 36468817

Number of columns in Movie Ratings table: 3
Number of rows in Movie Ratings table: 993153
```

PART 2 - Analyzing Genres

Let's now answer this question: how many unique genres are represented in this dataset?

Essentially, we have the genres per movie as a list - this is useful to quickly see what each movie might be represented as but it is difficult to easily answer questions such as:

- How many movies are categorized as Comedy, for instance?
- What are the top 20 most popular genres available?

Association Table

We need to "break out" these genres from the tconst? One common approach to take is to build an association table mapping a single tconst multiple times to each distinct genre.

For instance, given the following:

tconst	titleType	genres
abcd123	XXX	a,b,c

We would like to derive something like:

tconst	titleType	genre
abcd123	XXX	a
abcd123	XXX	b
abcd123	XXX	c

What this does is allow us to then perform a myriad of rollups and other analysis on this association table which can aid us in answering the questions asked above.

Implement the code necessary to derive the table described from the data set

In [16]:
```
genres.select("tconst","titleType","genres").show(5,truncate=False)
```

```
VBox()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+---------+---------+-----------------------+
|tconst   |titleType|genres                 |
+---------+---------+-----------------------+
|tt0000001|short    |Documentary,Short      |
|tt0000002|short    |Animation,Short        |
|tt0000003|short    |Animation,Comedy,Romance|
|tt0000004|short    |Animation,Short        |
|tt0000005|short    |Comedy,Short           |
+---------+---------+-----------------------+
only showing top 5 rows
```

Display the first 10 rows of your association table below

In [17]:
```
genres.withColumn("genre",explode(split("genres",","))).select("tconst","titleType","g
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+---------+---------+-----------+
|tconst   |titleType|genre      |
+---------+---------+-----------+
|tt0000001|short    |Documentary|
|tt0000001|short    |Short      |
|tt0000002|short    |Animation  |
|tt0000002|short    |Short      |
|tt0000003|short    |Animation  |
|tt0000003|short    |Comedy     |
|tt0000003|short    |Romance    |
|tt0000004|short    |Animation  |
|tt0000004|short    |Short      |
|tt0000005|short    |Comedy     |
+---------+---------+-----------+
only showing top 10 rows
```

Total Unique Genres

What is the total number of unique genres available in the movie category?

In [18]:
```
genres.withColumn("genre",explode(split("genres",",")))\
.filter((genres.titleType=="movie")&(genres.genres != '\\N') )\
.select("titleType","genre")\
.distinct()\
.count()



## This should not include genre=\N" so the result should be 28.- Thy Bui
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
28
```

What are the unique genres available?

In [19]:
```
genres.withColumn("genre",explode(split("genres",",")))\
.filter((genres.titleType=="movie")&(genres.genres != '\\N') )\
.select("genre")\
.distinct()\
.show(20)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+-----------+
|      genre|
+-----------+
|    Mystery|
|    Musical|
|     Action|
|      Sport|
|  Talk-Show|
|    Romance|
|   Thriller|
| Reality-TV|
|     Family|
|    Fantasy|
|    History|
|  Animation|
|  Film-Noir|
|      Short|
|     Sci-Fi|
|       News|
|      Drama|
|Documentary|
|    Western|
|     Comedy|
+-----------+
only showing top 20 rows
```

Oops! Something is off! OHH i actually got it in the previous code

In [20]:
```
genres.withColumn("genre",explode(split("genres",",")))\
.filter((genres.titleType=="movie")&(genres.genres != '\\N') )\
.select("genre")\
.distinct()\
.show(20)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

```
+-----------+
|      genre|
+-----------+
|    Mystery|
|    Musical|
|     Action|
|      Sport|
|  Talk-Show|
|    Romance|
|   Thriller|
| Reality-TV|
|     Family|
|    Fantasy|
|    History|
|  Animation|
|      Short|
|  Film-Noir|
|     Sci-Fi|
|       News|
|      Drama|
|Documentary|
|    Western|
|     Comedy|
+-----------+
only showing top 20 rows
```

In [21]:
```
genres.withColumn("genre",explode(split("genres",",")))\
.filter((genres.titleType=="movie")&(genres.genres != '\\N') )\
.select("titleType","genre")\
.distinct()\
.count()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
28
```

Top Genres by Movies

Now let's find the highest rated genres in this dataset by rolling up genres.

Average Rating / Genre

So now, let's unroll our distinct count a bit and display the per average rating value of per genre.

The expected output should be:

genre	averageRating
a	8.5
b	6.3
c	7.2

Or something to that effect.

First, let's join our two dataframes (movie ratings and genres) by tconst

In [24]:
```
genre_derived=genres.withColumn("genre",explode(split("genres",",")))\
.filter((genres.titleType=="movie")&(genres.genres != '\\N'))\
.select("tconst","titleType","genre")
join=genre_derived.join(movie_ratings,genre_derived.tconst==movie_ratings.tconst,"left
.sort(genre_derived.tconst.asc())\
.select("genre",col("averageRating").cast("float"))\
.show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+-----------+-------------+
|      genre|averageRating|
+-----------+-------------+
|    Romance|          5.4|
|Documentary|          5.2|
|      Sport|          5.2|
|       News|          5.2|
|      Drama|          6.2|
|  Biography|          6.2|
|  Biography|          6.1|
|      Crime|          6.1|
|      Drama|          6.1|
|      Drama|          4.8|
|      Drama|          2.7|
|      Drama|          4.2|
|      Drama|          3.6|
|    Fantasy|          4.8|
|  Adventure|          4.8|
|      Drama|          6.2|
|      Drama|          4.2|
|      Drama|          5.2|
|      Drama|          4.2|
|     Comedy|          4.0|
+-----------+-------------+
only showing top 20 rows
```

Now, let's aggregate along the averageRating column to get a resultant dataframe that displays average rating per genre.

In [25]:
```
genre_derived\
.join(movie_ratings,genre_derived.tconst==movie_ratings.tconst,"left")\
.sort(genre_derived.tconst.desc())\
.select("genre",col("averageRating").cast("float"))\
.groupBy("genre").agg(avg("averageRating").alias("avg_rating"))\
.show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

```
+-----------+-----------------+
|      genre|       avg_rating|
+-----------+-----------------+
|    Mystery|5.940437537126316|
|    Musical|6.203246053185319|
|     Action|5.718734067904495|
|      Sport|6.600145190943391|
|  Talk-Show|5.800000190734863|
|    Romance|6.125714179294426|
|   Thriller|5.625967567519544|
| Reality-TV|6.379310377712907|
|     Family|6.250560452699635|
|    Fantasy|5.924820762891499|
|    History|6.822718117193864|
|  Animation|6.326203749467441|
|      Short|7.259999942779541|
|  Film-Noir|6.636246780503378|
|     Sci-Fi|5.325150006900168|
|       News|7.200916040944689|
|      Drama|6.288080211097538|
|Documentary|7.245469805371099|
|    Western|5.948970991005059|
|     Comedy|5.941363107822231|
+-----------+-----------------+
only showing top 20 rows
```

Horizontal Bar Chart of Top Genres

With this data available, let us now build a barchart of all genres

HINT: don't forget about the matplotlib magic!

```
%matplot plt
```

In [26]:
```
genre_derived.join(movie_ratings,genre_derived.tconst==movie_ratings.tconst,"left")\
.select("genre",col("averageRating").cast("float"))\
.groupBy("genre").agg(avg("averageRating").alias("avg_rating"))\
.sort(col("avg_rating")\
      .desc())\
.show()
```

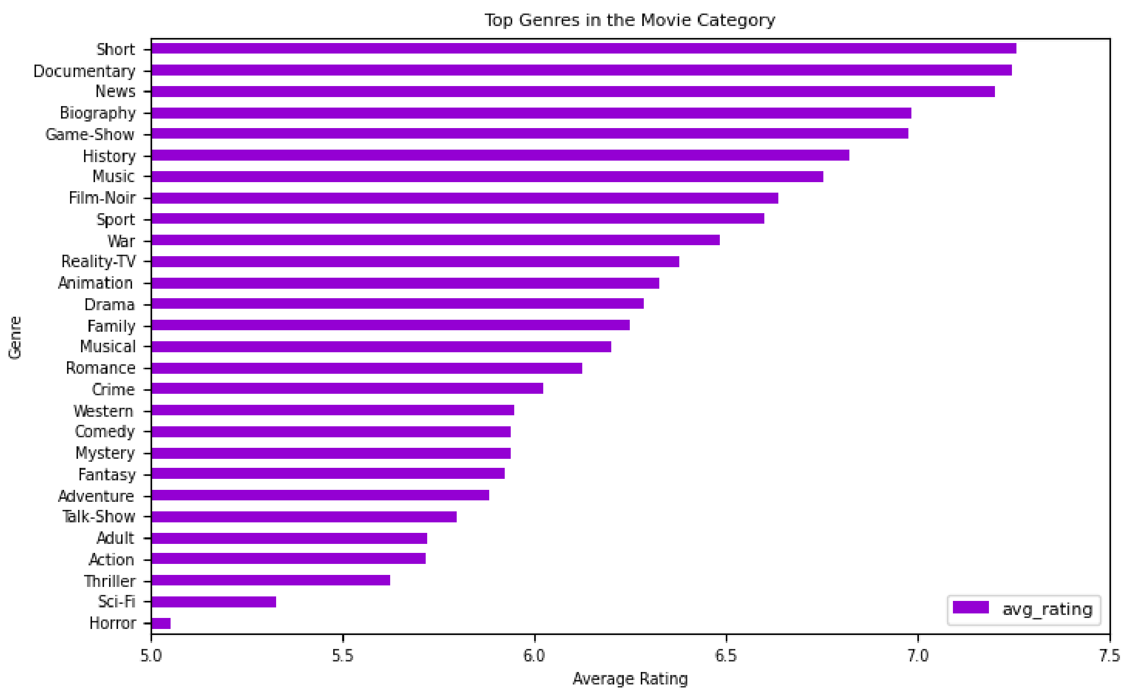
```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

```
+-----------+-----------------+
|      genre|       avg_rating|
+-----------+-----------------+
|      Short|7.259999942779541|
|Documentary|7.245469805371099|
|       News|7.200916040944689|
|  Biography|6.983637643044585|
|  Game-Show|6.974999904632568|
|    History|6.822718117193864|
|      Music|6.752020207214588|
|  Film-Noir|6.636246780503378|
|      Sport|6.600145190943391|
|        War|6.483807036278403|
| Reality-TV|6.379310377712907|
|  Animation|6.326203749467441|
|      Drama|6.288080211097538|
|     Family|6.250560452699635|
|    Musical|6.203246053185319|
|    Romance|6.125714179294426|
|      Crime|6.026013333109149|
|    Western|5.948970991005059|
|     Comedy|5.941363107822231|
|    Mystery|5.940437537126316|
+-----------+-----------------+
only showing top 20 rows
```

In [27]:

```
avg=genre_derived.join(movie_ratings,genre_derived.tconst==movie_ratings.tconst,"left"
.select("genre",col("averageRating").cast("float"))\
.groupBy("genre").agg(avg("averageRating").alias("avg_rating"))\
.sort(col("avg_rating")\
      .desc())
pdf=avg.toPandas().sort_values('avg_rating')
ax=pdf.plot(kind='barh',y='avg_rating',x='genre',color="#9400D3",fontsize="7",  figsiz
ax.set_xlabel("Average Rating", fontsize=7,fontname="Times New Roman")
ax.set_ylabel("Genre",fontsize=7,fontname="Times New Roman")
ax.set_title("Top Genres in the Movie Category",fontsize=8,fontname="Times New Roman")
plt.legend(prop={'size': 8})
%matplot plt
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```



PART 3 - Analyzing Job Categories

Total Unique Job Categories

What is the total number of unique job categories?

In [28]:
```
movie_actors\
.select("tconst","category")\
.distinct()\
.sort(col("tconst")\
      .asc())\
.show(5)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+---------+---------------+
|   tconst|       category|
+---------+---------------+
|tt0000001|       director|
|tt0000001|           self|
|tt0000001|cinematographer|
|tt0000002|       director|
|tt0000002|       composer|
+---------+---------------+
only showing top 5 rows
```

In [29]:
```
movie_actors.select("category").distinct().count()
```

```
VBox()
```

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
12
```

What are the unique job categories available?

In [30]:
```
movie_actors.select("category").distinct().show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+--------------------+
|            category|
+--------------------+
|             actress|
|            producer|
|production_designer|
|              writer|
|               actor|
|     cinematographer|
|       archive_sound|
|     archive_footage|
|                self|
|              editor|
|            composer|
|            director|
+--------------------+
```

Top Job Categories

Now let's find the top job categories in this dataset by rolling up categories.

Counts of Titles / Job Category

The expected output should be:

category	count
a	15
b	2
c	45

Or something to that effect.

In [31]:
```
movie_actors.groupBy("category").count().show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

```
+-------------------+-------+
|           category|  count|
+-------------------+-------+
|            actress|6325097|
|           producer|2197866|
|production_designer| 285924|
|             writer|4811596|
|              actor|8493701|
|     cinematographer|1300404|
|       archive_sound|   2143|
|     archive_footage| 209035|
|               self|6153089|
|             editor|1197669|
|           composer|1313187|
|           director|4179106|
+-------------------+-------+
```

Bar Chart of Top Job Categories

With this data available, let us now build a barchart of the top 5 categories.

HINT: don't forget about the matplotlib magic!

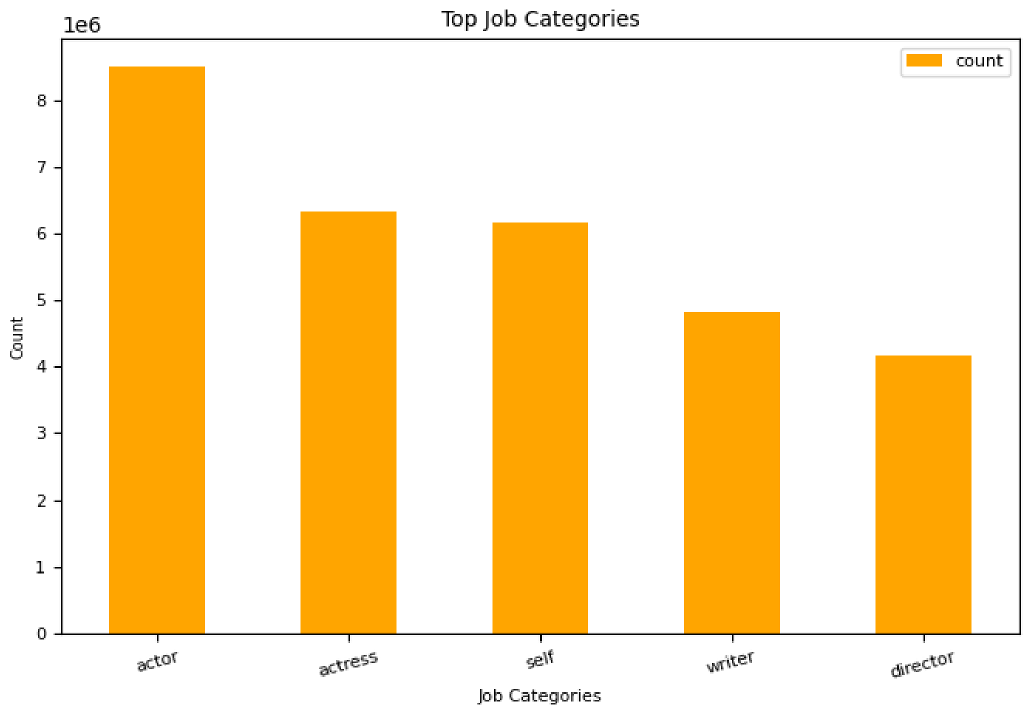
```
%matplot plt
```

In [32]:
```
movie_actors.groupBy("category").count().sort(col("count").desc()).show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+-------------------+-------+
|           category|  count|
+-------------------+-------+
|              actor|8493701|
|            actress|6325097|
|               self|6153089|
|             writer|4811596|
|           director|4179106|
|           producer|2197866|
|           composer|1313187|
|     cinematographer|1300404|
|             editor|1197669|
|production_designer| 285924|
|     archive_footage| 209035|
|       archive_sound|   2143|
+-------------------+-------+
```

In [33]:
```
count_cat=movie_actors.groupBy("category").count().sort(col("count").desc())
pdf3=count_cat.toPandas()
ax=pdf3.loc[0:4,:].plot(kind='bar',y='count',x='category',color="#FFA500",fontsize="8"
ax.set_xlabel("Job Categories", fontsize=8,fontname="Times New Roman")
ax.set_ylabel("Count",fontsize=7,fontname="Times New Roman")
ax.set_title("Top Job Categories",fontsize=10,fontname="Times New Roman")
plt.legend(prop={'size': 8})
plt.xticks(rotation=15)
%matplot plt
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```



PART 4 - Answer to the following questions:

In [35]:
```
## Prepare dataframe, filtered only Movie
tblMovie=movie_actors.select("tconst","nconst","category")
tblActors=actors.withColumn("title",explode(split("knownForTitles",","))).select("ncon
tblGenres=genres.filter((genres.titleType=="movie")).select("tconst","primaryTitle","s
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

1) Find all the "movies" featuring "Johnny Depp" and "Helena Bonham Carter".

First join actors, genres, and movie actors on each other

In [39]:
```
jointable=tblMovie\
.join(tblActors,on='nconst',how='left')\
.join(tblGenres,on='tconst',how='left')\
.select('primaryName','primaryTitle')\
.distinct()\
.sort(col('primaryTitle').asc())
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

In [40]:
```
## Filter list Johnny Depp and Helena Bonham Carter
filteractor=['Johnny Depp','Helena Bonham Carter']
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

In [41]:
```
## dataframe filtered out the 2 actors
tbl=jointable.filter(col('primaryName').isin(filteractor)).sort(col('primaryTitle').as
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

In [42]:
```
## Utilize sql to create two dataframes: one filtered Johnny and his movie, the other
tbl.createOrReplaceTempView("tbl")
JD=spark.sql("""
SELECT *
FROM tbl
WHERE primaryName ='Johnny Depp'
""")

HBC=spark.sql("""
SELECT *
FROM tbl
WHERE primaryName ='Helena Bonham Carter'
""")
## Find the movies JD and Helena starred
JD.join(HBC, on='primaryTitle',how='inner').select('primaryTitle').show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+--------------------+
|        primaryTitle|
+--------------------+
|Alice Through the...|
| Alice in Wonderland|
|Charlie and the C...|
|         Corpse Bride|
|        Dark Shadows|
|Sweeney Todd: The...|
+--------------------+
```

2) Find all the "movies" featuring "Brad Pitt" after 2010.

In [37]:
```
jointable1=tblMovie\
.join(tblActors,on='nconst',how='left')\
.join(tblGenres,on='tconst',how='left')\
.select("primaryTitle","startYear")\
.filter((col("primaryName")=="Brad Pitt")&(col('startYear')>2010))\
.distinct()\
.sort(col('startYear')\
      .desc())
jointable1.show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+--------------------+---------+
|        primaryTitle|startYear|
+--------------------+---------+
|             Babylon|     2021|
|        Irresistible|     2020|
|       Kajillionaire|     2020|
|            Ad Astra|     2019|
|            The King|     2019|
|Once Upon a Time ...|     2019|
|                Vice|     2018|
|         War Machine|     2017|
|Voyage of Time: L...|     2016|
|              Allied|     2016|
|       The Big Short|     2015|
|          By the Sea|     2015|
|    Hitting the Apex|     2015|
|                Fury|     2014|
|    12 Years a Slave|     2013|
|          Kick-Ass 2|     2013|
|         World War Z|     2013|
| Killing Them Softly|     2012|
|           Moneyball|     2011|
|    The Tree of Life|     2011|
+--------------------+---------+
```

3) What is the number of "movies" "acted" by "Zendaya" per year?

In [36]:
```
jointable2=tblMovie\
.join(tblActors,on='nconst',how='left')\
.join(tblGenres,on='tconst',how='left')\
.select("primaryTitle","startYear")\
.filter((col("primaryName")=="Zendaya")&(col("startYear")!= '\\N'))\
.distinct()\
.sort(col('startYear')\
      .desc())
jointable2.groupBy("startYear").count().show(truncate=True)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+---------+-----+
|startYear|count|
+---------+-----+
|     2020|    1|
|     2018|    2|
|     2017|    1|
+---------+-----+
```

4) What are the "movies" by average rating greater than "9.7" and released in "2019"?

In [38]:
```
movie_ratings.join(genres,on='tconst',how='left')\
.select("primaryTitle",col("averageRating").cast("float"))\
.filter((col('titleType')=="movie")&((col('startYear'))=='2019'))\
.sort(col('averageRating').desc()).distinct().show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
+--------------------+-------------+
|        primaryTitle|averageRating|
+--------------------+-------------+
|      A Grunt's Life|         10.0|
| The Butcher Baronet|         10.0|
|A Medicine for th...|         10.0|
|     Love in Kilnerry|        10.0|
|Bu Can Var Oldugu...|         10.0|
|   L'Enfant Terrible|         10.0|
|              Kirket|         10.0|
|   Our Scripted Life|         10.0|
|The Twilight Zone...|         10.0|
|          Superhombre|         9.9|
|         The Cardinal|          9.9|
|Puritan: All of L...|          9.9|
|Kamen Rider Zi-O:...|          9.8|
|     Time and motion|          9.8|
|We Shall Not Die Now|          9.8|
|   From Shock to Awe|          9.8|
|            Randhawa|          9.8|
|   Gini Helida Kathe|          9.8|
|          Square One|          9.8|
|         Freie Räume|          9.7|
+--------------------+-------------+
only showing top 20 rows
```

Extra Credit - Who "played" themself the most?

Try and analyze some interesting dimension to this data. You should specify the question in your Project2_Analysis.ipynb.

You must join at least two datasets.

In [48]:
```
self=movie_actors\
.filter(col("category")=='self')\
.join(actors,on='nconst',how='left')\
.select('primaryName')\
.groupBy('primaryName').count()

self.sort(col('count').desc()).show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(hei
ght='25px', width='50%'),…
```

```
+-----------------+-----+
|      primaryName|count|
+-----------------+-----+
|   Johnny Gilbert| 8520|
|      Alex Trebek| 7990|
|        Pat Sajak| 7071|
|       Bob Barker| 6948|
|     Johnny Olson| 6895|
|      Vanna White| 6838|
|       Ed McMahon| 6732|
|  David Letterman| 6171|
|       Dick Clark| 5959|
|  Carol Vorderman| 5723|
|Janice Pennington| 5501|
|         Jay Leno| 5237|
|        Gene Wood| 5222|
|    Johnny Carson| 5211|
|   Doc Severinsen| 5156|
|Charlie O'Donnell| 4731|
|     Paul Shaffer| 4677|
| Richard Whiteley| 4600|
|     Mike Douglas| 4423|
|             null| 4355|
+-----------------+-----+
only showing top 20 rows
```

I was wondering who is John Gelbert and what he does. Turned out he is a TV host. All of these personalities are TV hosts/comedians. Fun fact: John Gelbert is 94! No wonder he got the top results, with 8520 shows across his career.

In [ ]:

```
```